Multichromosomal Genome Median and Halving Problems

Eric Tannier[1], Chunfang Zheng[2], and David Sankoff[2]

[1] INRIA Rhône-Alpes, Université de Lyon 1, Villeurbanne, France
[2] University of Ottawa, Canada

Abstract. Genome median and halving problems aim at reconstructing ancestral genomes and the evolutionary events leading from the ancestor to extant species. The NP-hardness of the breakpoint distance and reversal distance median problems for unichromosomal genomes do not easily imply the same for the multichromosomal case. Here we find the complexity of several genome median and halving problems, including a surprising polynomial result for the breakpoint median and guided halving problems in genomes with circular and linear chromosomes; the multichromosomal problem is thus easier than the unichromosomal one.

1 Introduction

The gene order or syntenic arrangement of ancestral genomes may be reconstructed based on comparative evidence from present-day genomes — the phylogenetic approach — or on internal evidence in the case of genomes descended from an ancestral polyploidisation event, or from a combination of the two. The computational problem at the heart of phylogenetic analysis is the *median problem*, while internal reconstruction inspires the *halving problem*, and the combined approach gives rise to *guided halving*. How these problems are formulated depends first on the karyotypic framework: the number of chromosomes in a genome and whether they are constrained to be linear, and second on the objective function used to evaluate solutions. This function is based on some notion of genomic distance, either the number of *breakpoints*, adjacent elements on a chromosome in one genome that are disrupted in another, or the number of evolutionary operations necessary to transform one genome to another.

While the karyotypes allowed in an ancestor vary only according to the dimensions of single versus multiple chromosome, and linear versus circular versus mixed, the genomic distances of interest have proliferated according to the kinds of evolutionary operations considered, from the classic, relatively constrained, reversals/transocations distance to the more inclusive *double cut and join* measure, and many others.

The complexity of each of the problems is known for one or more distances, in one or more specific karyotypic contexts, and it is sometimes taken for granted that these results carry over to other combinations of context and distance. This is not necessarily the case. In this paper, we survey the known results and

K.A. Crandall and J. Lagergren (Eds.): WABI 2008, LNBI 5251, pp. 1–13, 2008.

unsolved cases for three distance measures in three kinds of karyotype, including several results presented here for the first time, including both new polynomial-time algorithms and NP-hardness proofs.

2 Genomes, Breakpoints and Rearrangements

Multichromosomal Genomes. We follow the general formulation of a genome in [3]. A *gene* A is an oriented sequence of DNA, identified by its *tail* A_t and its *head* A_h. Tails and heads are the *extremities* of the genes. An *adjacency* is an unordered pair of gene extremities; a *genome* is a set of adjacencies on a set of genes. Each adjacency in a genome means that two gene extremities are consecutive on the DNA molecule. In a genome, each gene extremity is adjacent to zero or one other extremity. An extremity x that is not adjacent to any other extremity is called a *telomere*, and can be written as an adjacency $x\circ$ with a null symbol $\circ$. Consider the graph G_Π whose vertices are all the extremities of the genes, and the edges include all the adjacencies in a genome Π as well as an edge joining the head and the tail of each gene. This graph is a set of disjoint paths and cycles. Every connected component is called a *chromosome* of Π. A chromosome is *linear* if it is a path, and *circular* if it is a cycle.

A genome with only one chromosome is called *unichromosomal*. These are *signed permutations* (linear or circular). A genome with only linear chromosomes is called a *linear* genome.

Genomes can be represented as a set of strings, by writing the genes for each chromosome in the order in which they appear in the paths and cycles of the graph G_Π, with a bar over the gene if the gene is read from the head to the tail (we say it has *negative* sign), and none if it is read from the tail to the head (it has *positive* sign). For each linear chromosome, there are two possible equivalent strings, according to the arbitrary chosen starting point. One is obtained from the other by reversing the order and switching the signs of all the genes. For circular chromosomes, there are also two possible circular string representations, according to the direction in which the cycle is traversed.

For example, if a genome Π is defined as the set of adjacencies on the set of genes $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

$$\{\circ 2_h, 2_t 1_h, 1_t 9_h, 9_t T, T 10_t, 10_h 6_h, 6_t 4_t, 4_h 3_h, 3_t T, T 8_t, 8_h 5_t, 5_h 7_t, 7_h \circ\},$$

we write it as the linear genome with 3 chromosomes:

$$\Pi = (\ \overline{2}\ \ \overline{1}\ \ \overline{9}\ ,\quad 10\ \ \overline{6}\ \ 4\ \ \overline{3}\ ,\quad 8\ \ 5\ \ 7\ )$$

A *duplicated gene* A is a couple of homologous oriented sequences of DNA, identified by two tails $A1_t$ and $A2_t$, and two heads $A1_h$ and $A2_h$. An *all-duplicates genome* Δ is a set of adjacencies on a set of duplicated genes.

For example, the following genome Δ is an all-duplicates genome on the set of genes $\{1, 2, 3, 4, 5\}$.

$$\Delta = (\ \overline{2}\ \ \overline{1}\ \ 2\ \ \overline{5}\ ,\quad 4\ \ \overline{3}\ \ 4\ \ \overline{1}\ ,\quad 3\ \ 5\ )$$

For a genome Π on a gene set $\mathcal{G}$, a *doubled genome* $\Pi \oplus \Pi$ is an all-duplicates genome on the set of duplicated genes from $\mathcal{G}$ such that if $A_x B_y$ is an adjacency of Π ($x, y \in \{t, h\}$), either $A1_x B1_y$ and $A2_x B2_y$, or $A2_x B1_y$ and $A1_x B2_y$ are adjacencies of $\Pi \oplus \Pi$. This includes telomeric adjacencies, so that a telomere in Π should yield two telomeres in $\Pi \oplus \Pi$.

Note the difference between a general all-duplicates genome and the special case of a doubled genome: the former has two copies of each gene, while in the latter these copies are organised in such a way that there are two identical copies of each chromosome (when we ignore the 1's and 2's in the $A1_x$'s and $A2_x$'s): it has two linear copies of each linear chromosome, and for each circular chromosome, either two circular copies or one circular chromosome containing the two successive copies. Note also that for a genome Π, there is an exponential number of possible doubled genomes $\Pi \oplus \Pi$ (exactly two to the power of the number of non telomeric adjacencies).

In discussing all-duplicates genomes, we will sometimes contrast them with *ordinary genomes* which have a single copy of each gene.

The Breakpoint Distance. We construct a breakpoint distance on multichromosomal genomes that depends on common adjacencies, or rather their absence, and also on common telomeres (or lack thereof) in two genomes. For two genomes Π and Γ on n genes, suppose Π has N_Π chromosomes, and Γ has N_Γ chromosomes. Let a be the number of common adjacencies, e be the number of common telomeres of Π and Γ. Then insofar as it should depend additively on these components, we may suppose the breakpoint distance has form

$$d_{BP}(\Pi, \Gamma) = n - a\beta - e\theta + (N_\Pi + N_\Gamma)\gamma + (|N_\Pi - N_\Gamma|)\psi,$$

where β, θ and γ are positive parameters, while ψ may have either sign. Taking $\Pi = \Gamma$ and imposing $d_{BP}(\Pi, \Pi) = 0$ yields the relations $\beta = 1$ and $1 - 2\theta + 2\gamma = 0$, so $\theta = \gamma + 1/2$. Now it is most plausible to count a total of 1 breakpoint for a fusion or fission of linear chromosomes, which implies $\gamma = \psi = 0$, so the most natural choice of *breakpoint distance* between Π and Γ is

$$d_{BP}(\Pi, \Gamma) = n - a - \frac{e}{2}.$$

For an all-duplicates genome Δ and an ordinary genome Π, the *breakpoint distance* between Π and Δ is $d_{BP}(\Pi, \Delta) = \min_{\Pi \oplus \Pi} d_{BP}(\Pi \oplus \Pi, \Delta)$.

The Double-Cut-and-Join Distance. Given a genome Π, which is defined as a set of adjacencies, a double-cut-and-join (DCJ) is an operation ρ acting on two adjacencies pq and rs (possibly some of p, q, r, s are $\circ$ symbols and even an adjacency may be composed of two $\circ$ symbols). The DCJ operation replaces pq and rs either by pr and qs, or ps and qr.

A DCJ can reverse an interval of a genome (DCJs include reversals), and may also fission one chromosome into two, fusion two chromosomes into a single one, or achieve a reciprocal translocation between two chromosomes. Two consecutive

DCJ operations may result in a block interchange: two segments of a genome exchange their positions, which results in a transposition if the two intervals are contiguous in the permutation. DCJ is thus a very general framework. It was introduced by Yancopoulos *et al.* [22], as well as by Lin *et al.* in a special case [14], and has since been adopted by Bergeron *et al.* [3] and many others, and has also been called "2-break rearrangement" [2].

If Π and Γ are two genomes on n genes, the *DCJ distance* $d_{DCJ}(\Pi, \Gamma)$ is the minimum number of DCJ operations needed to transform Π into Γ.

For an all-duplicates genome Δ and an ordinary genome Π, the *DCJ distance* between Π and Δ is $d_{DCJ}(\Pi, \Delta) = \min_{\Pi \oplus \Pi} d_{DCJ}(\Pi \oplus \Pi, \Delta)$.

The Reversal/Translocation Distance. The reversal/translocation distance was introduced by Hannenhalli and Pevzner [11], and is equivalent to the DCJ distance constrained to linear genomes.

If Π is a linear genome, a *linear* DCJ operation is a DCJ operation on Π that results in a linear genome. This allows reversals, reciprocal translocations, and chromosome fusions, fissions, which are special cases of translocations. Other DCJs, that create temporary circular chromosomes, are not allowed. If Π and Γ are linear genomes, the *RT* distance between Π and Γ is the minimum number of linear DCJ operations that transform Π into Γ, and is noted $d_{RT}(\Pi, \Gamma)$.

3 Computational Problems

The classical literature on genome rearrangements aims at reconstructing the evolutionary events and ancestral configurations that explain the differences between extant genomes. The focus has been on the genomic distance, median and halving problems. More recently the doubled distance and guiding halving problems have also emerged as important. In each of the ensuing sections of this paper, these five problems are examined for a specific combination of distance d (breakpoint, DCJ or RT) and kind of multichromosomal karyotype.

1- Distance. Given two genomes Π, Γ, compute $d(\Pi, \Gamma)$. Once the distance is calculated, an additional problem in the cases of DCJ and RT is to reconstruct the rearrangement scenario, i.e., the events that differentiate the genomes.

2- Double Distance. Given an all-duplicates genome Δ and an ordinary genome Π, compute $d(\Delta, \Pi)$. Because the assignment of labels "1" or "2" to the two identical (for our purposes) copies of a duplicated gene in Δ is arbitrary, the double distance problem is equivalent to finding such an assignment that minimises the distance between Δ and a genome $\Pi \oplus \Pi$ considered as ordinary genomes, where all the genes on any one chromosome in $\Pi \oplus \Pi$ are uniformly labeled "1" or "2" [2,26]. The double distance function is not symmetric because Δ is an all-duplicates genome and Π is an ordinary one, thus capturing the presumed asymmetric temporal and evolutionary relationship between the ancestor Π and the present-day genome Δ.

3- Median. Given three genomes Π_1, Π_2, Π_3, find a genome M which minimises $d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M)$. The median problem estimates the common ancestor of two genomes, given a third one (not necessarely specifies) as an outgroup. It may be used as a hint for phylogenetic studies.

4- Halving. Given an all-duplicates genome Δ, find an ordinary genome Π which minimises $d(\Delta, \Pi)$. The goal of a halving analysis is to reconstruct the ancestor of an all-duplicates genome at the time of the doubling event.

5- Guided Halving. Given an all-duplicates genome Δ and an ordinary genome Π, find an ordinary genome M which minimises $d(\Delta, M) + d(M, \Pi)$. The guided halving problem is similar to the genome halving problem for Δ, but it takes into account the ordinary genome Π of an organism presumed to share a common ancestor with M, the reconstructed undoubled ancestor of Δ.

We will survey these five computational problems for the three distances that we have introduced, in the cases of multichromosomal genomes containing all linear chromosomes or permitting circular chromosomes.

4 Breakpoint Distance, General Case

In this section, $d = d_{BP}$, and genomes are considered in their most general definition, that is, multichromosomal with both circular and linear chromosomes allowed. As the nuclear genome of a eukaryotic species, such a configuration would be rare and unstable. Nevertheless this case is of great theoretical interest, as it is the only combination of distance and karyotype where all five problems mentioned in Section 3 prove to be polynomially solvable, including the only genomic median problem that is polynomially solvable to date. Furthermore, the solutions in this context may suggest approaches for other variants or the problems, as well as providing a rapid bound for other distances, through the Watterson *et al.* bound [21].

Distance and Double Distance. The distance computation follows directly from the definition, and is easily achievable in linear time.

The double distance computation is also easy: let Π be a genome and Δ be an all-duplicates genome. Let ab be an adjacency in Π (a or b may be $\circ$ symbols). If $a1b1$ or $a2b2$ is an adjacency in Δ, choose $a1b1$ and $a2b2$ for adjacencies in $\Pi \oplus \Pi$. If $a1b2$ or $a2b1$ is an adjacency in Δ, choose $a1b2$ and $a2b1$ for adjacencies in $\Pi \oplus \Pi$. The two cases are mutually exclusive, so the assignment is made without ambiguity. Assign all remaining adjacencies arbitrarily.

It is easy to see that this procedure minimises $d(\Pi \oplus \Pi, \Delta)$, as every possible common adjacency or telomere in Δ and Π is a common adjacency or telomere in $\Pi \oplus \Pi$ and Δ.

Median. The following result contrasts with the NP-completeness proofs of all genome median problems in the literature [6,7,17]. The problem is NP-complete

for unichromosomal genomes, whether they are linear or circular [6,17], but the multichromosomal case happens to be easier.

Theorem 1. *There is a polynomial time algorithm for the multichromosomal genome median problem.*

Proof. For this extended abstract, we show only the principle of the algorithm, and omit the details of the proofs. Let Π_1, Π_2, Π_3 be three genomes on a gene set $\mathcal{G}$ of size n. Draw a complete graph G on the vertex set containing the union of all the extremities of the genes in $\mathcal{G}$ and a set containing one supplementary vertex t_x for every gene extremity x. For any pair of gene extremities x, y, weight the edge xy by the number of genomes, among Π_1, Π_2, Π_3, for which xy is an edge. Then each edge in G joining two gene extremities is weighted by 0, 1, 2 or 3. Now for any vertex x, weight the edge xt_x by half the number of genomes, among Π_1, Π_2, Π_3, having x as a telomere. Each edge xt_x is then weighted by 0, $\frac{1}{2}$, 1, or $\frac{3}{2}$. To every other edge of the complete graph G, assign the weight 0.

Let M be a perfect matching in G. Clearly, the edges between gene extremities in G define the adjacencies of a genome, that we also call M. The relation between the weight of the perfect matching M and the median score of the genome M is easy to state:

Claim. The weight of the perfect matching M in G is $3n-(d(\Pi_1, M)+d(\Pi_2, M)+d(\Pi_3, M))$.

This implies that a maximum weight perfect matching M is a minimum score median genome. As the maximum weight perfect matching problem is polynomial, so is the breakpoint median problem. □

Note that this algorithm remains valid if the median of more than three genomes is to be computed.

Halving. To our knowledge, the genome halving with breakpoint distance has not yet been studied. In this framework, it has an easy solution, using a combination of elements from the maximum weight perfect matching technique in Theorem 1 and the double distance computation: let Δ be an all-duplicates genome on a gene set $\mathcal{G}$, and G be the complete graph on the vertex set containing all the extremities of the genes in $\mathcal{G}$, plus one supplementary vertex t_x for every gene extremity x. For any pair of gene extremities x, y, weight the edge xy by zero, one or two, according to the number of times an xy adjacency is present in Δ. Now for any vertex x, weight the edge xt_x by half the number of times x is a telomere in Δ. Weight the remaining edges between t vertices by zero.

Claim. A maximum weight perfect matching M in G defines the adjacencies of a genome M minimising $d(\Delta, M)$.

Guided Halving. Again, this will be the only polynomial result for the guided genome halving problem. The solution combines elements of the three algorithms (double distance, median, halving) previously discussed in this section.

Let Δ be an all-duplicates genome on a gene set $\mathcal{G}$, and Π be a genome on $\mathcal{G}$. Let G be the complete graph on the vertex set containing all the extremities of the genes in $\mathcal{G}$, plus one supplementary vertex t_x for every gene extremity x. For any pair of gene extremities x, y, weight the edge xy by the number of times x is adjacent to y in Δ and Π, and weight the edge xt_x by half the number of times x is a telomere in Δ and Π. Weight the remaining edges between t vertices by zero.

Claim. A maximum weight perfect matching M in G defines the adjacencies of a genome M minimising $d(\Delta, M) + d(M, \Pi)$.

5 Breakpoint Distance, Linear Case

In this section, $d = d_{BP}$ and all genomes must be linear, as is most appropriate for modeling for the eukaryotic nuclear genome. The solutions to the distance and double distance problems are the same as in the previous section, where circularity was allowed. But in contrast to the model of Section 4, all the problems concerning at least three genomes are NP-complete.

The Median Problem

Theorem 2. *The breakpoint median problem for multichromosomal linear genomes is NP-hard.*

Proof. For this extended abstract, we show only the principle of the reduction, and omit the details. We use a reduction from the circular permutation median (CPM) problem, which asks: Given three circular genomes Π_1, Π_2, and Π_3 with only one chromosome, find a circular genome M with only one chromosome, which minimises $d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M)$. This problem is NP-hard [6,17].

Let Π_1, Π_2, Π_3, be an instance of the CPM problem, on the gene set $\{1, \ldots, n\}$. Let $n+1$ be a new gene, and Π_i' be the genome constructed from Π_i $(1 \leq i \leq 3)$ by deleting the adjacency $x1_t$ (x is the extremity of a gene in $\{2, \ldots, n\}$), and adding the adjacency $x(n+1)_t$. Genomes Π_1', Π_2' and Π_3' are linear. Let k be a positive integer.

Claim. There exists a unichromosomal and circular genome M on $\{1, \ldots, n\}$ with $d(\Pi_1, M)+d(\Pi_2, M)+d(\Pi_3, M) \leq k$ if and only if there exists a multichromosomal and linear genome M' on $\{1, \ldots, n+1\}$ with $d(\Pi_1', M') + d(\Pi_2', M') + d(\Pi_3', M') \leq k$. (This claim implies the theorem). □

Halving and Guided Halving. Surprisingly, the halving problem has not been treated in the literature. We conjecture it has a polynomial solution, because the halving problem for all other rearrangement distances is polynomial. Constructing a solution is beyond the scope of this paper, and the problem remains open.

This guided halving problem is NP-hard, as proved in [24], using the NP-completeness result for the median proved above.

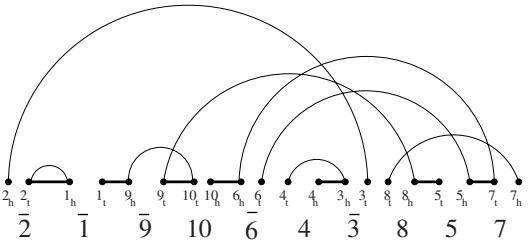

Fig. 1. The breakpoint graph of the genomes $\Pi = (\overline{2}\ \overline{1}\ \overline{9}\ 10\ \overline{6}\ ;\ 4\ \overline{3}\ ;\ 8\ 5\ 7)$ and $\Gamma = (1\ 2\ 3\ 4\ ;\ 5\ 6\ 7\ 8\ 9\ 10)$. Π-edges are drawn with bold segments, and Γ-edges are the thin circle arcs.

6 DCJ Distance, General Case

In this section, $d = d_{DCJ}$. Genomes can have several chromosomes, circular or linear. This is the most general context in which the DCJ distance has been explicitly formulated [3].

The complexity of the genome median probem is not established by the work of Caprara [7], who proved the unichromosomal result only. We show its NP-hardness here. The double distance problem was proposed by Alekseyev and Pevzner [2], and we will show its NP-hardness as well.

Distance. There is an easy linear solution, both for the distance and the scenario computation [3,22].

The *breakpoint graph* of two genomes Π and Γ, denoted by $BP(\Pi, \Gamma)$, is the bipartite graph whose vertex set is the set of extremities of the genes, and there is an edge between two vertices x and y if xy is an adjacency in either Π (these are Π-edges) or Γ (Γ-edges). Note that we do not invoke any $\circ$ symbols. Vertices in this graph have degree zero, one or two, so that the graph is a set of paths (possibly including some with no edges) and cycles. It is also the line-graph of the *adjacency graph*, an alternate representation in [3], and is commonly used in genome rearrangement studies. Fig. 1 shows an example of a breakpoint graph. Theorem 3 shows how to obtain the distance directly from the graph.

Theorem 3. *[3][1] For two genomes Π and Γ, let $c(\Pi, \Gamma)$ be the number of cycles of the breakpoint graph $BP(\Pi, \Gamma)$, and $p(\Pi, \Gamma)$ be the number of paths with an even number of edges. Then*

$$d(\Pi, \Gamma) = n - c(\Pi, \Gamma) - \frac{p(\Pi, \Gamma)}{2}.$$

Note the similarity to the breakpoint distance formula in Section 2. The number of genes n is the same in both formulae, the parameter c is related to parameter

[1] The formula is presented in [3] with the cycles and odd paths of the adjacency graph. This corresponds to cycles and even paths of the breakpoint graph, as it is the line-graph of the adjacency graph.

a in the breakpoint formula in that each common adjacency is a cycle of the breakpoint graph (with two parallel edges), and parameter p is related to parameter e, as each shared telomere is an even path (with no edge) in the breakpoint graph. Although these two measures of genomic distance were derived in different contexts and through different reasoning, their formulae show a remarkably similar form. They differ in that the DCJ formula counts non trivial cycles and paths, but for distant genomes, they tend to give the same values.

Double Distance

Theorem 4. *The DCJ double distance problem is NP-hard for multichromosomal genomes.*

Proof. For this extended abstract, we show only the principle of the reduction, and omit the details. Reduction is from the *breakpoint graph decomposition* (BGD) problem (see [7]). A graph G is *bicoloured* if all its edges are coloured in either red or blue; it is *balanced* if it has only degree 2 or degree 4 vertices, every vertex is incident to the same number of red and blue edges, and there is no cycle formed by only red or only blue edges.

Given a balanced bicoloured graph G, the breakpoint graph decomposition problem is to find a decomposition of the edges of G into a maximum number of edge-disjoint cycles, each alternating between red and blue edges. Berman and Karpinski [4] proved APX-hardness of this problem.

Let G be a balanced bicoloured graph on n vertices, defining an instance of the BGD problem. Define the gene set $\mathcal{G}$ as the vertex set of G. Construct an all-duplicates genome Δ and a genome Π on $\mathcal{G}$ in the following way. First, for each vertex X of G, let X_t and X_h be its extremities; let X_tX_h be an adjacency in Π. Then, for every vertex X of G, let $X1_t$, $X1_h$, $X2_t$ and $X2_h$, be the extremities of the duplicated gene X. For each blue edge XY in G, construct an adjacency in Δ joining the heads of genes $X1$ or $X2$, and $Y1$ or $Y2$: if vertex X has degree 4, one of the two adjacencies defined by the two blue edges involves $X1_h$, and the other $X2_h$ (arbitrarily). If vertex X has degree 2, define the adjacency with $X1_h$ and add another adjacency $X1_tX2_h$ in Δ. For each red edge, add an adjacency in Δ according to the same principle, but joining tails of genes.

We then have an all-duplicates genome Δ, and a genome Π. Note that Π is composed of n circular chromosomes, one for each gene, and that neither Π nor Δ have telomeres.

Claim. The maximum number of edge-disjoint alternating cycles in G is equal to $2n - d(\Delta, \Pi)$. (This claim implies the theorem). □

Median. Though effective heuristics are available [1], we have:

Theorem 5. *The DCJ median problem for multichromosomal genomes is NP-hard.*

Proof. We use a reduction from the breakpoint graph decomposition defined in the proof of Theorem 4, in a way very similar to part of Caprara's proof [7] for the unichromosomal case.

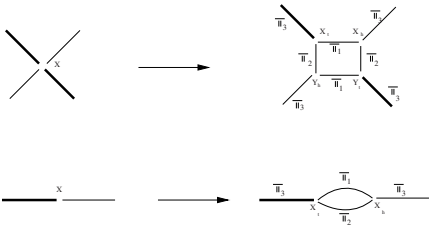

Fig. 2. Strategy for reducing the breakpoint graph decomposition to the DCJ median problem. Red edges are represented by thick lines, while blue edges are thin.

Let G be a balanced bicoloured graph on n vertices. Define the gene set $\mathcal{G}$ as a set containing one gene X for every degree 2 vertex of G, and two genes X and Y for every degree 4 vertex of G.

Apply the following transformation to G, which is similar to the transformation in [7], as illustrated in Figure 2.

Let v be a vertex of degree 2 in G. Replace v by two vertices labeled by the two extremities of the associated gene X, namely X_t and X_h. The blue edge incident to v becomes incident to X_h and the red edge to X_t. Add one Π_1-edge and one Π_2-edge between X_h and X_t. Now let v be a vertex of degree 4 in G. Replace v by four vertices labeled by the four extremities of X and Y, X_t, X_h, Y_h, and Y_t. The blue edges incident to v become incident to X_h and Y_h, while the red edges become incident to X_t and Y_t. Add two Π_1-edges X_tX_h and Y_tY_h, and two Π_2-edges X_tY_h and Y_tX_h. Red and blue edges are the Π_3-edges. Call the final graph G'.

It is easy to see that Π_1, Π_2, and Π_3 define genomes on the set of genes $\mathcal{G}$, and they have no telomeres. Let $w2$ be the number of degree 2 vertices of Γ, and $w4$ be the number of degree 4 vertices of Γ.

Claim. There exists a genome M such that $d(M,\Pi_1)+d(M,\Pi_2)+d(M,\Pi_3) \leq w2+3w4-k$ if and only if there exists at least k edge-disjoint alternating cycles in G. (This claim implies the theorem.) □

Halving and Guided Halving. This problem has a polynomial solution, as recently stated for unichromosomal genomes by [2] and in the general case by [15,20]. All these algorithms are simplified versions of the algorithm by El-Mabrouk and Sankoff [9], developed for the RT rearrangement distance.

Theorem 6. *Guided halving is NP-complete for multichromosomal genomes.*

We omit here the proof of this theorem, based on a reduction of the same problem and similar ideas than in the previous one.

7 DCJ and Reversal/Translocation, Linear Chromosomes

In the original formulation of the DCJ distance [22], it was shown that there is a solution where each excision of a circular intermediate could be followed

directly by its reinsertion. Thus the median and halving problems can be stated in terms of exclusively linear chromosomes in both the data genomes and the reconstructed ancestor. They all remain open.

Hannenhalli and Pevzner proposed a polynomial-time algorithm for calculating $d_{RT}(\Pi, \Gamma)$ for two genomes Π and Γ [11]. This was reformulated in [19] and minor corrections were added by [16] and [12]. A polynomial time genome halving algorithm was given in [9]. Though the constrained DCJ distance in the preceding paragraph is arguably just as realistic, because of the long history of d_{RT}, effective heuristics have been developed and applied for the double distance [23,26], median [5,13] and guided halving problems [23,25,26], but their complexities remain open. Note that Chen *et al.* give an NP-completeness result on a problem which slightly generalizes the RT double-distance probem.

8 Conclusions

Table 1. Current knowledge of the status of complexity questions for five problems related to ancestral genome reconstruction, for eight genomic distances in the unichromosomal and multichromosomal contexts, including the new results in this paper. Other versions of the halving problem are less restrictive [2,9,20]. P and NP stand for polynomial and NP-hard, respectively; when followed by ?, represent our conjectures.

problem context	distance	halving	double distance	median	guided halving
breakpoint uni	P	open	P	NP [6,17]	open
breakpoint general multi	P new	P new	P new	P new	P new
breakpoint linear multi	P new	open P?	P new	NP new	NP [24]
DCJ uni	P [3,22]	P [2]	open	NP [7]	open
DCJ general multi	P [3,22]	P [15,20]	NP new	NP new	NP new
DCJ linear multi	P [22]	open	open	open NP?	open NP?
RT uni	P [10]	open	open	NP [7]	open
RT multi	P [11,12,16,19]	P [9]	open NP?	open NP?	open NP?

Acknowledgements

Research supported in part by a grant to DS and a doctoral fellowship to CZ from the Natural Sciences and Engineering Research Council of Canada (NSERC). ET is funded by the Agence Nationale pour la Recherche (GIP ANR JC05_49162 and NT05-3_45205) and the Centre National de la Recherche Scientifique (CNRS). DS holds the Canada Research Chair in Mathematical Genomics.

References

1. Adam, Z., Sankoff, D.: The ABCs of MGR with DCJ. Evol. Bioinform. 4, 69–74 (2008)
2. Alekseyev, M., Pevzner, P.: Colored de Bruijn graphs and the genome halving problem. TCBB 4, 98–107 (2008)

3. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Bücher, P., Moret, B.M.E. (eds.) WABI 2006. LNCS (LNBI), vol. 4175, pp. 163–173. Springer, Heidelberg (2006)
4. Berman, P., Karpinski, M.: On some tighter inapproximability results: Further improvements. ECCC Report 65, Univ. of Trier. (1998), http://www.eccc.uni-trier.de/
5. Bourque, G., Pevzner, P.A.: Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Res. 12, 26–36 (2002)
6. Bryant, D.: The complexity of the breakpoint median problem. TR CRM-2579. Centre de recherches mathématiques, Université de Montréal (1998)
7. Caprara, A.: The reversal median problem. INFORMS J. Comput. 15, 93–113 (2003)
8. Chen, X., Zheng, J., Fu, Z., Nan, P., Zhong, Y., Lonardi, S., Jiang, T.: Assignement of orthologous genes via genome rearrangement. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2, 302–315 (2005)
9. El-Mabrouk, N., Sankoff, D.: The reconstruction of doubled genomes. SIAM J. Comput. 32, 754–792 (2003)
10. Hannenhalli, S., Pevzner, P.A.: Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. JACM 46, 1–27 (1999)
11. Hannenhalli, S., Pevzner, P.A.: Transforming men into mice (polynomial algorithm for genomic distance problem). In: FOCS 1995, pp. 581–592 (1995)
12. Jean, G., Nikolski, M.: Genome rearrangements: a correct algorithm for optimal capping. Inf. Process Lett. 104, 14–20 (2007)
13. Lenne, R., Solnon, C., Stützle, T., Tannier, E., Birattari, M.: Reactive stochastic local search algorithms for the genomic median problem. In: van Hemert, J., Cotta, C. (eds.) EvoCOP 2008. LNCS, vol. 4972, pp. 266–276. Springer, Heidelberg (2008)
14. Lin, Y.C., Lu, C.L., Chang, H.Y., Tang, C.: An efficient algorithm for sorting by block-interchange and its application to the evolution of vibrio species. JCB 12, 102–112 (2005)
15. Mixtacki, J.: Genome halving under DCJ revisited. In: Hu, X., Wang, J. (eds.) COCOON 2008. LNCS, vol. 5092. Springer, Heidelberg (2008)
16. Ozery-Flato, M., Shamir, R.: Two notes on genome rearrangement. JBCB 1, 71–94 (2003)
17. Pe'er, I., Shamir, R.: The median problems for breakpoints are NP-complete. In: ECCC (1998)
18. Sankoff, D., Blanchette, M.: The median problem for breakpoints in comparative genomics. In: Jiang, T., Lee, D.T. (eds.) COCOON 1997. LNCS, vol. 1276, pp. 251–263. Springer, Heidelberg (1997)
19. Tesler, G.: Efficient algorithms for multichromosomal genome rearrangements. JCSS 65, 587–609 (2002)
20. Warren, R., Sankoff, D.: Genome halving with general operations. APBC 2008, Adv. Bioinform. Comput. Biol. 6, 231–240 (2008)
21. Watterson, G., Ewens, W., Hall, T., Morgan, A.: The chromosome inversion problem. J. Theoret. Biol. 99, 1–7 (1982)
22. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinform. 21, 3340–3346 (2005)
23. Zheng, C., Wall, P.K., Leebens-Mack, J., de Pamphilis, C., Albert, V.A., Sankoff, D.: The effect of massive gene loss following whole genome duplication on the algorithmic reconstruction of the ancestral Populus diploid. In: CSB 2008 (in press, 2008)

24. Zheng, Z., Zhu, Q., Adam, Z., Sankoff, D.: Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. In: ISMB 2008 (in press, 2008)
25. Zheng, Z., Zhu, Q., Sankoff, D.: Genome halving with an outgroup. Evol. Bioinform. 2, 319–326 (2006)
26. Zheng, Z., Zhu, Q., Sankoff, D.: Descendants of whole genome duplication within gene order phylogeny. In: JCB (in press, 2008)